1 A primer on molecular biology

Alexander Zien

Modern molecular biology provides a rich source of challenging machine learning problems. This tutorial chapter aims to provide the necessary biological background knowledge required to communicate with biologists and to understand and properly formalize a number of most interesting problems in this application domain.

The largest part of the chapter (its first section) is devoted to the cell as the basic unit of life. Four aspects of cells are reviewed in sequence: (1) the molecules that cells make use of (above all, proteins, RNA, and DNA); (2) the spatial organization of cells ("compartmentalization"); (3) the way cells produce proteins ("protein expression"); and (4) cellular communication and evolution (of cells and organisms). In the second section, an overview is provided of the most frequent measurement technologies, data types, and data sources. Finally, important open problems in the analysis of these data (bioinformatics challenges) are briefly outlined.

1.1 The Cell

The basic unit of all (biological) life is the cell. A *cell* is basically a watery solution of certain molecules, surrounded by a lipid (fat) membrane. Typical sizes of cells range from 1 μm (bacteria) to 100 μm (plant cells). The most important properties of a living cell (and, in fact, of life itself) are the following:

Life

- It consists of a set of molecules that is separated from the exterior (as a human being is separated from his or her surroundings).
- It has a metabolism, that is, it can take up nutrients and convert them into other molecules and usable energy. The cell uses nutrients to renew its constituents, to grow, and to drive its actions (just like a human does).
- It is able to (approximately) replicate, that is, produce offspring that resemble itself.
- It can react to its environment in a way that tends to prolong its own existence

and the existence of a (preferably high) number of offspring.

Viruses, which are simpler than cells, also satisfy some definitions that characterize life: they can, for example, reproduce. But because they depend so strongly on the help of host cells and they do not have their own metabolism, viruses are usually not considered to be alive.

Eukarya, prokarya

Two types of living organisms can be distinguished: *prokarya* (further subdivided into *eubacteria* and *archaea*), which are always single cells, and *eukarya* (which include all animals, plants, and fungi). Eukaryotic cells are more complex than prokarya in that their interior is more organized: the eukaryote is divided into so-called compartments. For instance, the *nucleus* contains hereditary information, and a number of *mitochondria* serve to supply the cell with certain energy-rich molecules.

The incredibly complex machinery of cells cannot be decently described in this short chapter. An excellent and detailed overview can be found in the textbook by Alberts et al. (2002), or, in a shortened version, in Alberts et al. (1998). Here, we try to provide some rough impressions of the absolute basics.

1.1.1 Important Molecules of the Cell

Cells are defined by the molecules they are composed of. Especially important for the integrity of cells are three kinds of macromolecules, which are now introduced. These molecules are *polymers*, which means that they are composed of a large number of covalently[1] linked *monomers*, small molecular building blocks. The set of different monomers and the way they are linked determine the type of polymer.

Nucleotides

DNA The major part of the heritable information of a cell is stored in the form of DNA molecules. They are called the cell's *genome*. *DNA (deoxyribonucleic acid)* is a chain molecule that is composed of linearly linked nucleotides. *Nucleotides* are small chemical compounds. There are essentially four different nucleotides that occur in cellular DNA, which are usually called A (adenine), C (cytosine), G (guanine), and T (thymine).[2] The chain of nucleotides has a direction, because its two ends are chemically different. Consequently, each DNA molecule can be described by a text over a four-letter alphabet. Chemists denote its beginning as the *5′-end* and its end as the *3′-end*. The two directions are denoted by *upstream*, for "towards" the beginning, and *downstream*, for "towards" the end. Molecular chains of only a few nucleotides are called *oligonucleotides*.

[1] Among the different types of bonds that are possible between atoms, covalent bonds are the strongest. Molecules are defined as the smallest covalently connected sets of atoms; they are often represented by graphs of covalent connections.

[2] The restriction to four nucleotides is a simplification that is sufficient for most bioinformatics analysis. In reality, in genomic DNA cytosines may be methylated. This modification can be biologically significant, but it is usually not revealed in the available data.

DNA is a good carrier of information that is supposed to be retained for a long time (in fact, usually for the lifetime of a cell, which can be years). DNA can form very stable structures due to the following properties. The nucleotides A and T can bind to each other by forming two hydrogen bonds; therefore, A and T are said to be *complementary.* G and C are also complementary: they form three hydrogen bonds. Importantly, the ability to bind in this way holds for chains of nucleotides, that is, for DNA molecules. The *complement* of a DNA sequence is the sequence of the complements of its bases, but read in the reverse direction; complements are often called *complementary DNA (cDNA).* Complementary strands can bind to each other tightly by forming a double helix structure, which enables all the hydrogen bonds between the pairs of complementary bases. The binding of two complementary DNA molecules is often referred to as *hybridization.*

Complementarity, hybridization

In cells, the genomic DNA is indeed present in the form of a double helix of two complementary strands, as illustrated in figure 1.1. Apart from the increased stability, this provides redundancy, which serves the cell in two ways. First, erroneous changes from one nucleotide to another, termed *point mutations*, can thereby be detected and corrected. Second, there is a natural way to duplicate the genome, which is necessary when the cell divides to produce two daughter cells. The double helix is separated into two single strands of DNA, each of which then serves as a template for synthesizing its complement. Since the complement of a complement of a DNA sequence is again the primary sequence, the above procedure results in two faithful copies of the original double-stranded DNA.

The size of genomes can be enormous; for instance, the human genome consists of more than 3 billion nucleotides. Although the human genome is separated into 23 separate DNA molecules, each part still has an average length of about 5 cm —about 5000 times longer than the diameter of a human cell! Consequently, the DNA in cells is kept in a highly packaged form. In regular intervals, assemblies of proteins (called *histones*) bind to the DNA. The DNA double helix winds about one and a half times around each histone complex to form a *nucleosome*; the nucleosomes resemble beads on a string (of DNA). The nucleosomes themselves are usually packed on top of one another to form a more compact fibrous form called *chromatin.* An even higher level of packing is achieved by introducing loops into the chromatin fiber. The resulting structures, one for each genomic DNA molecule, are known as *chromosomes.* They do not flow around freely in the nucleus, but are anchored to nuclear structures at sites called *matrix attachment regions (MARs).*

Genome packing, chromosomes

In many organisms, two or more versions of the genome may be present in a cell. This is called a *diploid* or *polyploid* genome. In contrast, a single set of chromosomes is said to be *haploid.* In sexual organisms, most cells contain a diploid genome, where one version is inherited from each parent. The germ cells giving rise to offspring contain a haploid genome: for each chromosome, they randomly contain either the maternal or the paternal version (or a mixture thereof).

Ploidy

RNA *RNA (ribonucleic acid)* is very similar to DNA: again, it consists of nucleotides linked in a chain. In contrast to DNA, the nucleotide U (for uracil) is used

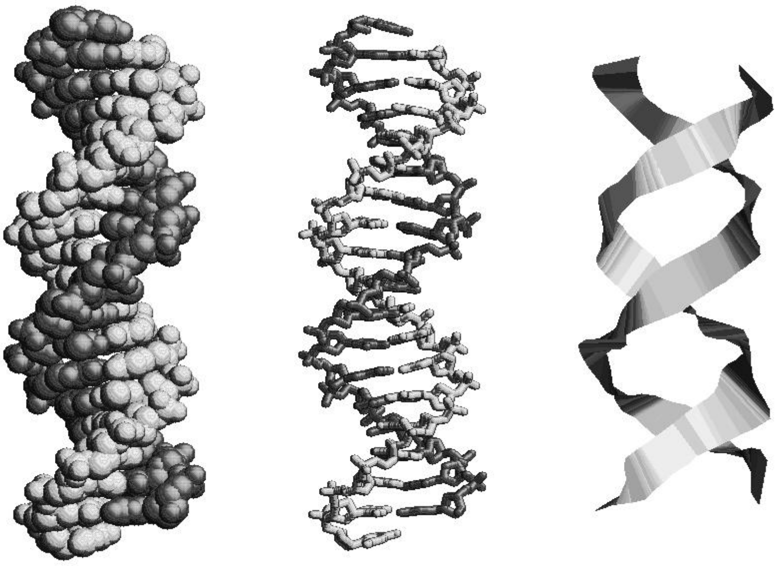

Figure 1.1 The double helix structure of genomic DNA. The same piece of DNA is visualized (using the program RasMol) in three different ways of increasing abstraction. *Left*, spacefill: each atom is shown as a ball. *Middle*, covalent bonds between heavy atoms are shown as sticks. *Right*, each strand of the double helix is shown as a ribbon. (DNA part of PDB entry `1hcq`.)

instead of T, and the chemical details of the nucleotides differ slightly. Due to these difference RNA molecules are usually single-stranded, which allows them to form a variety of structures in three-dimensional (3D) space that can perform complex tasks (such RNAs are called *ribozymes*).

Genes

The importance of the genome is that it typically contains many genes. Although there is still debate about the exact definition, a *gene* can be thought of as a substring of the genome that is responsible for the production of one or a couple of types of RNA molecules. In the process of *gene expression*, the RNA is synthesized to be complementary to a part of the DNA template. As a result, each gene can control one or more properties of the organism, although often quite indirectly, as will become apparent below.

Note that genes also include parts of DNA that are not copied into RNA. Most important, each gene contains a sequence called a promoter, which specifies the conditions under which RNA copies of certain parts of the gene are produced. Although ribozymes are responsible for a few very important tasks in cells, the purpose of the vast majority of genes in a cell is to encode building instructions for proteins (certain macromolecules; see the next paragraph). The RNA molecules involved in this process are called *messenger RNAs*, or *mRNAs*. In figure 1.2, the flow of information from the DNA to the proteins is illustrated.

mRNA

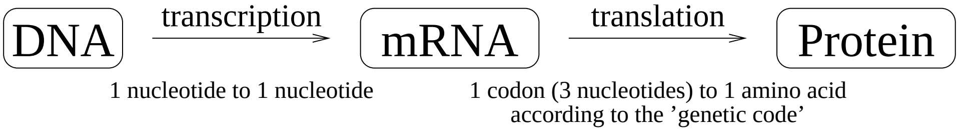
DNA
transcription
mRNA
translation
Protein
1 nucleotide to 1 nucleotide
1 codon (3 nucleotides) to 1 amino acid according to the 'genetic code'

Figure 1.2 Flow of genetic information.

Amino acids

Proteins *Proteins* are polymers composed of amino acids. Cells use 20 different types of amino acids for protein synthesis. Common to each amino acid are two chemical groups (an amino [N] group and a carboxyl [C] group) which form *peptide bonds* (a special kind of covalent bond) to link two amino acids. Since a water molecule is split off during the formation of such a bond, a protein is actually composed of *amino acid residues* (often, just *residues*). Proteins are also sometimes called *polypeptides* (most commonly in contexts where their 3D structures are not important); molecules consisting of only a few amino acids are called *oligopeptides*, or simply *peptides*. Due to their chemistry, the beginning and the end of a protein are called its *N-terminus* and its *C-terminus*, respectively. The chain of peptide links forms the *backbone* of a protein. Importantly, each amino acid also has a third group, the *side chain*. The side chains of the 20 natural amino acids show very different chemical properties.

Tertiary structure, fold

Each polypeptide *folds* into an elaborate spatial structure, called its *tertiary structure* or, sloppily, its *fold*. Figure 1.3 should convey an impression of the typical complexity of a fold by showing this structure for an arbitrary protein (in two common graphical representations). The tertiary structure depends on the particular sequence of amino acids (which is also sometimes called the *primary structure* and can be represented by a text over a 20-letter alphabet). The 3D structure of natural proteins[3] is usually assumed to be uniquely determined by the sequence (given cellular conditions such as acidity, the concentrations of ions, etc.).[4] However, sometimes the cell must help to achieve this uniqueness. In these cases, other proteins, named *chaperones*, guide the folding process.

Secondary structure

Among the *structural motifs*, i.e., spatial structures of subpeptides occurring in proteins, two are of exceptional importance: the α helix and the β strand. In an *α helix*, consecutive amino acids assume the shape of a spiral with 3.6 amino acids per turn. This motif is especially stable due to a regular pattern of weak bonds between any amino acid and the fourth next amino acid. In a β strand, the backbone is extended and can gain stability from neighboring β strands. The resulting structure, a *β-sheet*, again shows a regular pattern of weak bonds, this time between the strands. Together with the *coils*, which subsume the remainder of the protein, α helices and β strands are the elements of *secondary structure*. This

[3] That is, proteins that occur in some natural life form.

[4] Some studies suggest that the majority of possible sequences do not fold into a unique structure. However, nature appears to prefer sequences that do so; presumably, well-defined structures can contribute more to the cell's survival.

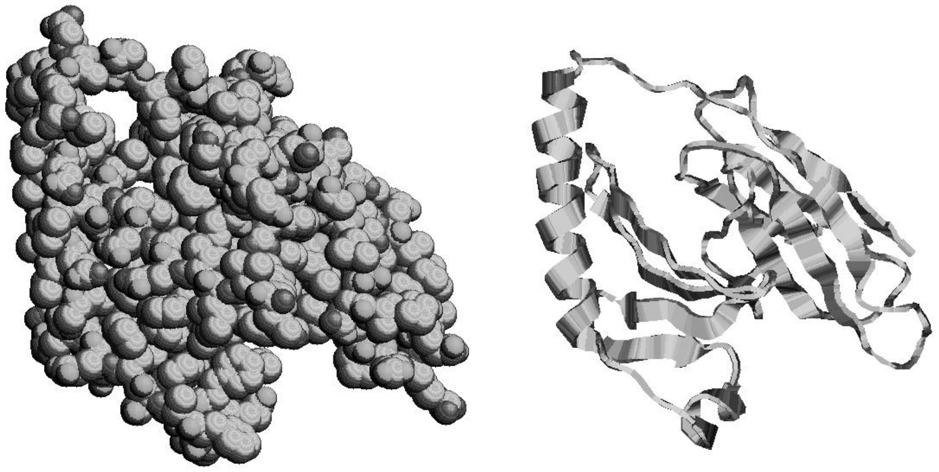

Figure 1.3 Two visual representations of the same protein. *Left*, full atom view. *Right*, the more useful cartoon view. (From PDB entry `1A6A`, drawn by RasMol.)

is often exploited in schematic representations of proteins, as illustrated in figure 1.4. The secondary structure already determines in large part the complete protein fold, the *tertiary structure*.

Domains

A *domain* is a subunit of a protein which can fold separately into its native structure. Especially in higher organisms (multicellular eukaryotes), many multidomain proteins have evolved, supposedly because recombining domains is an efficient way of creating new proteins that perform useful functions.

Binding

The structure of the backbone of a folded protein determines its overall shape, and also which amino acids are exposed on the surface. Due to the diversity of the side chains, this allows for the generation of a huge variety of patterns of physicochemical properties on protein surfaces. The surface properties determine which other molecules the protein can bind to. The (cellular) function of a protein can most immediately be defined by its set of binding partners and the chemical reactions induced by the binding. For example, many proteins that bind small molecules have cavities, called *binding pockets*, into which the *ligand* (the specific small molecule) fits like a key into a lock. Frequently, the function of a protein requires it to bind to two or more other molecules. This is often achieved through a separate domain for each binding partner.

Functions

The functions of proteins in cells are as diverse as the tasks that cells have to perform. Functional categories include (but are not limited to) the following:

- *Metabolism.* Proteins called *enzymes* bind small molecules called *metabolites* to catalyze reactions yielding other small molecules. In this way, nucleotides for DNA and RNA, amino acids for proteins, lipids for membranes, and many other essential compounds are produced. Cells may be viewed as tiny but highly complex and

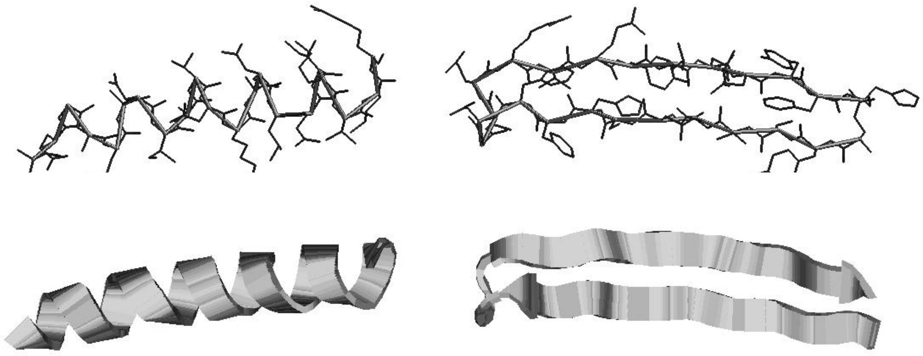

Figure 1.4 Secondary structure elements. *Left*, an α helix. *Right*, a β sheet with two strands. *Top*, stick model of the covalent bonds between heavy atoms; the backbone is emphasized by the thicker sticks. *Bottom*, cartoon view. (From PDB entry `1A6A`, visualized with RasMol.)

competent chemical factories.

- *Energy.* This can be seen as a special case of metabolism, because cells produce a few types of small molecules as energy carriers.
- *Transcription, protein synthesis, and protein processing.* The huge machinery required to produce proper proteins from DNA is, to a great extent, run by proteins (although ribozymes play a crucial role, too).
- *Transport and motor proteins.* Cells can be more efficient due to a nonrandom spatial distribution of molecules. In particular, compartmentalized cells contain elaborate transport mechanisms to achieve and maintain appropriate local concentrations. Molecular motion can even become visible on a macroscopic scale: muscle contractions are driven by the motion of myosin proteins on actin filaments (longish intracellular structures built from actin proteins).
- *Communication (intra- or intercellular).* Communication is most important for multicellular organisms. While signaling molecules are usually much smaller than proteins, they are received and recognized by proteins. The processing of signals allows computations to be performed; this may be most obvious for the human brain (involving $\sim 10^{11}$ cells), but also underlies the directed motion of unicellular organisms.
- *Cell cycle.* Most cells (be they alone or part of a multicellular organism) recurrently divide into two daughter cells to reproduce. This complex process is orchestrated and carried out by proteins.

A complete list of protein functions is far beyond the scope of this chapter. In summary, proteins are major building blocks of the cell and, above all, the machines that keep cells running.

Macromolecules We have now met the three most important types of macromolecules in the cell (DNA, RNA, and protein) and their relation (the genetic flow of information). A fourth type of macromolecule which also occurs in cells shall only briefly be mentioned here: the polysaccharide. *Polysaccharides* are polymers composed of covalently linked *monosaccharides* (sugars, such as glucose, fructose, galactose). In contrast to the macromolecules discussed earlier, their bonding pattern is not necessarily linear, but often rather treelike. Examples illustrating the relevance of polysaccharides are starch, which is the principal food reserve of plants; glycogen, the counterpart of starch in animals; cellulose, a major constituent of the cell walls of plants; and chitin, which makes up most of the exoskeleton of insects.

Saccharides

In table 1.1, all four types of macromolecules and their most important properties and functions are summarized. Table 1.2 shows their contributions to the total mass of a cell, also in comparison to smaller types of molecules to be described below; not surprisingly, proteins dominate.

Proteins, RNA, and DNA can be parts of even more intricate *assemblies* or, synonymously, *complexes*. For example, as described above, histone proteins are used to pack DNA into chromatin. The *ribosome*, which performs the translation of mRNAs to proteins, is a huge assembly of several proteins and ribosomal RNA (rRNA). The individual molecules in an assembly (which are not connected by covalent bonds) are referred to as *subunits*. Just to make things more confusing, (stable) complexes of proteins (in the sense of individual translation products, as introduced above) are sometimes also called *proteins*; the subunits are then also called *(protein) chains*.

Complexes

Membrane *Membrane* is another huge assembly of smaller units. It mainly consists of a bilayer of lipids (of several different types). A membrane is not a macromolecule, because the lipids are not covalently connected (i.e., they remain separate molecules). Instead, the lipids stick together because they are largely *hydrophobic*, which means that they repel water. By forming a bilayer, all hydrophobic parts contact other hydrophobic surfaces (of other lipid molecules). Only the *hydrophilic* (water-loving) heads of the longish lipids face the water.

Hydrophobicity

The hydrophobicity of the membrane interior prevents water and molecules dissolved in water (which are hydrophilic) from penetrating the membrane. Thus, a membrane is used to separate the cell from its exterior: no large or hydrophilic compounds can pass it directly. In eukaryotes, membranes also serve to enclose compartments (which are subspaces of the cell with distinct chemical properties). To admit the controlled exchange of molecules and also of information, membranes also contain many proteins (often in the sense of protein complexes) that stick out on both sides. The surface of such *membrane-proteins* typically features a hydrophobic ring where it is embedded into the membrane.

Metabolites Of course, small molecules are vital for cells, too. Here we give just a few selected examples:

Macro-molecule	**DNA**	**RNA**
Building blocks	nucleotides (A,C,G,T)	nucleotides (A,C,G,U)
Typical length	1000s to 10^9s	100s to 1000s
Structure	double helix, tightly packed and organized in several levels	complex 3D structure, with structural motifs (secondary structure)
Function	storage of (most of) the hereditary information of an organism: the genome, which contains the genes as subsequences	■ *messenger RNA (mRNA)*: serves as the blueprint for protein production ■ *transfer RNA (tRNA)*: connects codons to amino acids (implementing the genetic code); used by the ribosome ■ *ribosomal RNA (rRNA)*: forms part of the ribosome (amounting to ∼90% of the total RNA)
Location	nucleus, mitochondria, chloroplasts	nucleus, cytosol, mitochondria, chloroplasts

Macro-molecule	**Protein**	**Polysaccharides**
Building blocks	amino acids (20 different types)	monosaccharides (several types)
Typical length	10s to 1000s	up to 10^9 (e.g., starch)
Structure	complex and versatile, with structural motifs (secondary structure, domains, etc.)	often not linearly bonded but treelike
Function	Extremely diverse. For example, ■ *enzymes* catalyze reactions of other molecules; ■ *structural proteins* build and stabilize the structure of the cell; ■ *receptors*, *kinases*, and other proteins receive, transport, and process signals from the exterior; ■ *transcription factors (TF)* regulate the production of all proteins.	■ modification of proteins and their properties ■ storage of energy (e.g., in starch) ■ structural stability (e.g., in chitin) ■ storage of water (e.g., in extracellular matrix in cartilage)
Location	everywhere in- and outside cell; dissolved in water or embedded in a membrane	everywhere in- and outside cell; often bound to proteins

Table 1.1 Important macromolecules of the cell. They are composed of small molecules, covalently connected to form linear chains.

Molecule	**Cell Mass in**	
type	**Bacteria**	**Mammals**
H_2O (water)	70%	70%
DNA	1%	0.25%
RNA	6%	1%
proteins	15%	18%
lipids (fat)	2%	5%
polysaccharides (sugar)	2%	2%
metabolites and inorganic ions	4%	4%

Table 1.2 Approximate fractions of different classes of molecules of the total weight of a typical cell.

- Adenosine triphosphate (ATP) and NADPH (both derived from the nucleotide A) serve as ubiquitous ready-to-use sources of energy.
- Monosaccharides (sugars) and lipids (fats) can be converted into ATP, and therefore serve as a long-term source of energy. Saccharides are also often attached to proteins to modify their properties.
- *Signaling molecules* convey information by docking to their respective receptor proteins and triggering their action. For example, steroids (which include many sex hormones) can diffuse into a cell's nucleus and induce the activation of some genes.

Small molecules are more generally called *compounds.*

1.1.2 Compartmentalization of the Eukaryotic Cell

As mentioned earlier, eukaryotic cells contain many *compartments*, which are also called *organelles.* They are subspaces that are enclosed by single or double membranes. Figure 1.5 provides an overview of the major compartments in the cell, and table 1.3 summarizes some of their properties.

In each compartment, a cell maintains different concentrations of relevant molecules. This way, the compartmentalization allows the cell to perform diverse tasks and chemical reactions that require different environments (e.g., a certain acidity) efficiently. As an example, the bulk of a cell's hereditary information is stored as DNA molecules (condensed into chromatin) in the nucleus, where the transcription machinery (which produces mRNA copies from genes) can more easily find it than if the DNA were allowed to reside anywhere in the cell.

Subcellular localization

Since each type of compartment is devoted to different tasks in the cell, each requires a distinct set of proteins to perform the subtasks. In order to save resources, proteins are specifically delivered to the organelles that require them. Consequently, many proteins contain signals that specify their destination. These signals can either be entire peptides (e.g., hydrophobic stretches for transfer into the endoplasmatic reticulum [ER]) or characteristic surface patches of the folded protein. There are

Compartment	**Function(s)**	**Membrane**
Cytosol	protein synthesis, general metabolism, etc.	single
Nucleus	■ storage of main genome (DNA molecules) ■ RNA synthesis ■ ribosome synthesis (in the nucleolus)	double
Endoplasmatic reticulum (ER) (inner space of nuclear membrane, extending throughout the cell)	■ synthesis of most lipids (membrane) ■ synthesis of proteins for single-membrane organelles (rough ER) ■ post-translational processing of those proteins	single
Golgi apparatus	■ post-translational processing of proteins ■ distribution of proteins and lipids to single-membrane organelles	single
Vesicles (mobile bubbles)	transport of proteins and membrane between single-membrane organelles and to/from cell exterior	single
Endosomes	■ contain material taken up from the exterior; or ■ secrete contents (mainly proteins) to cell exterior	single
Lysosomes/vacuoles (plants, fungi)	digest of molecules, organelles, etc. / store waste and nutrients, control cell size	single
Peroxisomes	carry out oxidative (dangerous) reactions	single
Cell exterior / extracellular matrix	■ extracellular matrix connects cells, stabilizes the organism, contains nutrients, etc. ■ in polarized cells (e.g., nerve cells), the exterior is divided into basolateral and apical parts	single
Mitochondria	generate ATP by oxidizing nutrients	double
Chloroplasts (in plants)	generate energy-rich molecules from sunlight	double

Table 1.3 Important compartments of the eukaryotic cell. Chloroplasts (which occur in plants, but not in animals) and mitochondria contain their own (small) genomes and produce a part of the proteins they require themselves; they have probably evolved from enclosed prokaryotic cells.

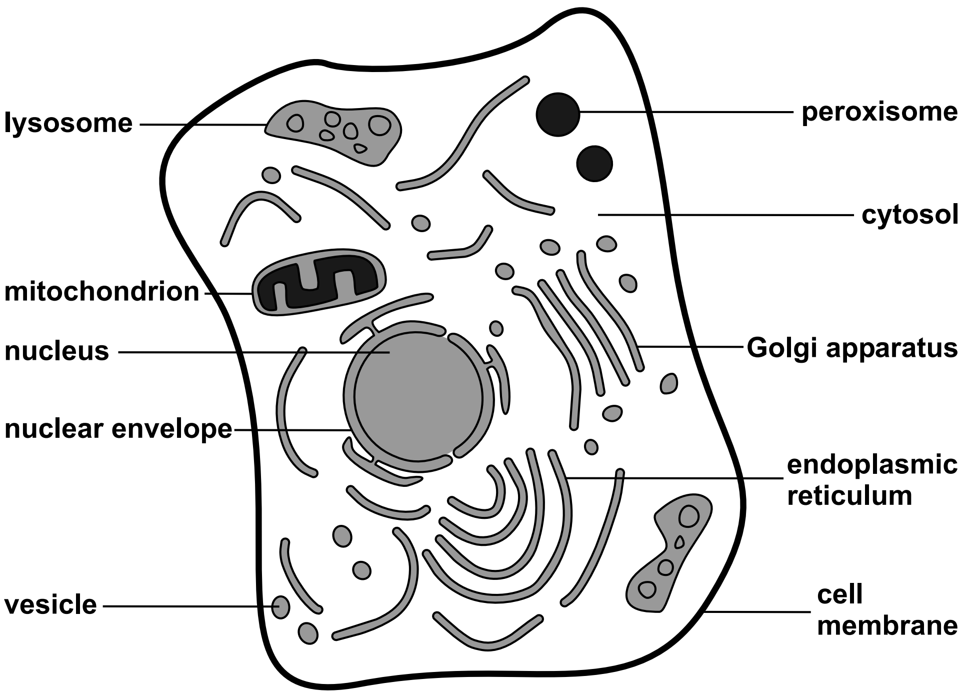

Figure 1.5 Compartments in a eukaryotic cell. All lines represent membranes. The interior of all compartments is shaded in gray; the cytosol is white. Inspired by a figure in Alberts et al. (1998) and crafted by Karin Bierig.

also default destinations when signals are absent: proteins showing no signal at all stay in the cytosol. The *subcellular localization* is obviously closely related to the function of the protein.

It should be noted that cells are in general not spatially symmetric. For example, the surface of many cells in multicellular organisms is divided into two domains: the apical and the basolateral. An extreme case is provided by nerve cells: their apical part consists of *axons*, thin extensions (that can be as long as 2 *m* in the human) which connect a neuron to other neurons. The *exocytotic* pathway, which transports proteins to the cell exterior, can distinguish between the two regions.

1.1.3 Expression of Genes and Proteins

One of the most fundamental processes in the cell is the production (and disposal) of proteins. Below, the life cycle of proteins is outlined for eukaryotic cells.

1. *Transcription. Messenger RNA (mRNA)* copies of a gene are produced. The products, called *pre-mRNAs* (since they are not yet spliced; see step 2), are complementary to the DNA sequence.

(a) Initiation: Certain proteins, called *transcription factors (TFs)*, bind to *TF* *binding sites* in the gene promoters in the DNA.

(b) Elongation: The mRNA copy of the gene is synthesized by a special protein (RNA polymerase II). It moves along the DNA and thereby sequentially extends the pre-mRNA by linking a nucleotide complementary to that found in the DNA.

(c) Termination: A signal in the DNA causes the transcription to end and the mRNA to be released.

2. *Splicing.* Parts of the pre-mRNA, which are called *introns*, are removed. The remaining parts, called *exons*, are reconnected to form the mature mRNA. The spliced mRNAs travel from the nucleus (through huge, selective pores in its double membrane) into the cytosol. To increase the chemical stability of the mRNA, a chemical cap is formed at the $5'$-end and a *poly(A)* sequence (built from many A nucleotides) is appended to the $3'$-end.

3. *Translation.* In the cytosol, ribosomes await the mRNAs. Ribosomes synthesize proteins as specified by *codons* —triplets of consecutive nucleotides—in the mRNA.

(a) Initiation: The ribosome finds a *start codon* (usually, the first AUG subsequence that has favorable neighboring nucleotides) in the mRNA.

(b) Elongation: One by one, the ribosome attaches amino acids to the growing polypeptide (protein) chain. In each step, the ribosome translates the current codon into an amino acid according to the *genetic code*. The ribosome then moves to the next codon in the same *reading frame*, that is, to the next adjacent nonoverlapping codon.

(c) Termination: Translation is stopped by any of three different *stop codons* encountered in the current reading frame.

4. *(Posttranslational) modification* (not for all proteins). The protein may be chemically modified, if it contains the relevant signals and if it resides in a compartment where these signals are recognized.

(a) Additional chemical groups can be covalently attached to proteins (glycosylation (sugars), phosphorylation, methylation, etc).

(b) Covalent bonds can be formed between amino acids.

(c) Proteins can be covalently bound to each other.

(d) Proteins can be cleaved, that is, cut into parts.

5. *Translocation* (not for all proteins). Proteins are delivered to the appropriate compartment, which is specified by signals in the amino acid sequence. The signal can either be a typical short segment of a sequence, or a structural motif on the surface of the protein (which may be composed of amino acids that are not neighbors in the sequence). In the absence of signals, the protein stays in the cytosol.

6. *Degradation.* Almost all proteins are eventually destroyed by digestion into their individual amino acids.

In prokaryotes, the entire process is a bit less complex because splicing is uncommon and the translocation has only three different targets (cytosol, membrane, exterior) due to the lack of compartments.

Alternative splicing

The process of splicing implies complex *gene structures* composed of alternating introns and exons; an illustration is given in figure 1.6. However, it allows for increased flexibility by a mechanism known as *alternative splicing*: certain proteins can cause certain exons to be lengthened, shortened, or even skipped completely. Thus, the same gene can give rise to the production of different proteins. This is an important way for cells to adapt to the circumstances, including their cell type and extracellular signals. It is estimated that a human gene on average encodes for eight or nine different proteins. More detailed information on the process of splicing and its biological implications can be found in chapter **??**, section **??**.

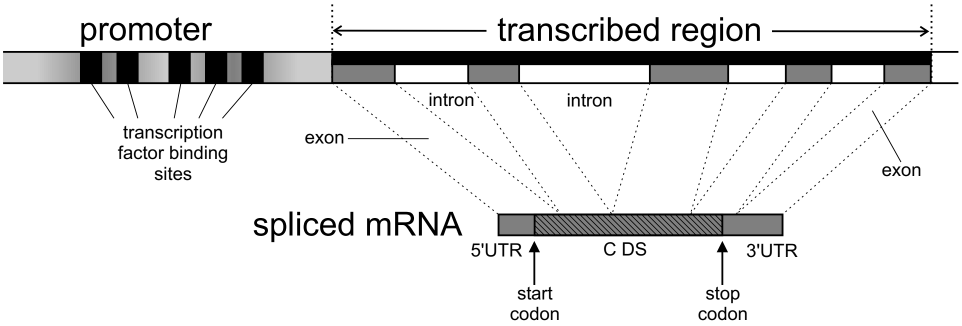

Figure 1.6 Typical gene structure in eukaryotes. At the top of the figure, a section of a genome is shown that contains a gene. Important features of the gene are the promoter containing several TF binding sites and the transcribed region, which is partitioned into exons and introns. After transcription, the pre-mRNA (not shown) is spliced: the introns are cut out; thus, the mature mRNA is the concatenation of the exons. Only a part of the mRNA encodes a protein (CDS, for coding sequence); the other parts are called UTRs (untranslated regions). Artwork by Karin Bierig.

Expression levels

Steps 1 and 2 of the scheme described above are called *gene expression*, while steps 1 through 5 are called *protein expression*. The term *expression level* of a molecule type is (a bit imprecisely) used to refer to either its current abundance in the cell, or to the rate of synthesis of new molecules. This difference is often neglected for gene expression, which may or may not be justified by the fact that mRNAs are degraded relatively quickly after having been translated several times. However, for proteins the distinction is crucial, because their lifetimes may be very long and differ vastly.

Regulation

The cellular concentration of any type of protein can be influenced by changing the efficiencies of the above steps. This is called *regulation* of expression. While cells in fact regulate each of the above steps, the main point for the quantitative control of protein expression is certainly transcription initiation. In addition to the general TFs, which are always required for initiation, there are additional TFs which modify the probability or speed of transcription. They bind to short DNA motifs, for obvious reasons called *enhancers* and *silencers*, in the promoter. The effects of TF binding sites can extend over huge distances in the DNA sequence; therefore *insulators* (certain DNA signals) may be required to separate genes from each other and prevent mutual regulatory interference.

The steps of protein expression have a natural temporal ordering, where each step operates on the result of the preceding step. However, there are at least three types of deviation from a clear, serial manufacturing process: (1) Some of the steps may occur concurrently, or can be performed before the preceding step is finished. For example, much of the splicing is carried out while the gene is still being transcribed. Also, the translocation from the cytosol into the ER and some modifications take place during translation. (2) There is no compulsory ordering of translocation and modification. In fact, many proteins are modified in the ER and the Golgi apparatus, which are intermediate stations on the journey to their destination compartment (cf. section 1.1.2). (3) Degradation may occur even before the protein is finished and delivered.

In many cases, the mentioned exceptions relate to the *folding* of the newly synthesized protein into a 3D structure. A protein can already start to fold while it is still growing out of the ribosome, and modifications by other proteins at that time can have an impact on the way it folds. Some proteins are aided in finding the desired structure by helper proteins (*chaperones*), which, for instance, unfold incorrectly folded proteins. In case a protein repeatedly misfolds (i.e., does not assume the intended structure despite the help of chaperones), it can also be degraded.

1.1.4 Beyond the Cell

Cell Communication Cells, especially those in the same multicellular organism, can communicate by the exchange of extracellular *signal molecules*. This way, the coordinated action of many (in the case of a grown human, on the order of 10^{10} or 10^{11}) cells can be achieved. Even to perform no action requires signaling; in animals, cells that do not get a constant supply of certain signals from their neighbors commit *apoptosis*, that is, self-destruction. This is a safety provision used to eliminate malfunctioning cells; if the mechanism gets broken itself, uncontrolled proliferation (cancer) may result.

Depending on the properties of the emitted signal molecules, the signaling can affect neighboring cells only (*contact-dependent*); it can be locally restricted to a small cluster of cells (*paracrine*); or it can rely on distribution through the blood system (*endocrine*). A special case is the *synaptic signaling*, in which the electric signal transmitted by a neuron causes neurotransmitters to be released that induce an electric potential in the receiving neuron. As in endocrine signaling, the signal is carried over large distances (as electrical potential through the long axons of the neurons). But in contrast to endocrine signaling, the signaling molecules travel only a very short distance extracellularly and are transmitted to a very specific set of target cells.

Extracellular signals can take one of two routes into the cell: through *cell surface receptors* or directly to *intracellular receptors.* For the latter, the signaling molecules have to traverse the cell membrane. This is possible for small hydrophobic molecules like steroid hormones (which include, e.g., the sex hormone testosterone). When such a molecule binds to the corresponding receptor protein, the protein usually travels to the nucleus (or it may already have been there) and activates or inhibits the transcription of one or several genes.

Signaling molecules that cannot permeate the cell membrane are recognized by cell surface receptors, to which they bind extracellularly. These receptors reside in the membrane and also have an intracellular part. The extracellular binding of the signaling molecule induces an action by the intracellular part, for example, a change of the 3D structure. In response to this change, a series of downstream actions begins: cytosolic proteins modify each other in a chain, until finally a TF is activated or deactivated or the reaction rate of an enzyme is altered.

Evolution The complexity of cells and organisms has evolved over several billion years of interplay of mutation and selection. Here, *mutation* means any kind of modification of the heritable information (basically the genome) of reproductive cells. The totality of heritable information giving rise to an organism is called its *genotype*, as opposed to *phenotype*, which subsumes the observable physical properties of the organism. Differences in the genotype sometimes manifest in different phenotypes; otherwise, the corresponding mutations are said to be *silent.*

Selection

Selection refers to the fact that the phenotypic changes may lead to differential reproductive success (e.g., some mutations are directly lethal); this may correlate with the organism's ability to survive in its environment. Often, however, mutations have no or a negligible impact on survival and reproduction (even if they are not silent). Several different genotypes (and possibly phenotypes) may then coexist in a population. In this case, their genetic differences are called *polymorphisms.*

Mutations

There are several different types of mutations. The simplest is the *point mutation* or *substitution*; here, a single nucleotide in the genome is changed. In the case of polymorphisms, they are called *single nucleotide polymorphisms (SNPs).* Other types of mutations include the following:

- *Insertion.* A piece of DNA is inserted into the genome at a certain position.
- *Deletion.* A piece of DNA is cut from the genome at a certain position.
- *Inversion.* A piece of DNA is cut, flipped around and then re-inserted, thereby converting it into its complement.
- *Translocation.* A piece of DNA is moved to a different position.
- *Duplication.* A copy of a piece of DNA is inserted into the genome.

The term *rearrangement* subsumes inversion and translocation.

While mutations can be detrimental to the affected individual, they can also in rare cases be beneficial; or, much more frequently, just neutral (under the

Genetic diversity

actual circumstances). Thereby mutations can increase the *genetic diversity* of a population, that is, the number of present polymorphisms. In combination with selection, this allow a species to adapt to changing environmental conditions and to survive in the long term. For example, many viruses (such as HIV) have imprecise replication mechanisms to produce a large fraction of mutants among the huge number of descendants. This way, subpopulations are created that do not match the patterns that the immune system of their host is looking for.

Many bacteria have evolved a strategy to achieve more complex mutations: by *horizontal gene transfer*, genetic material is not received from parental cells, but from other cells which may even belong to a different species. *Transposons*—mobile segments of DNA that can move around or copy themselves within the genome—presumably also serve to generate (certain kinds of) mutations. Sexual reproduction can be viewed as a sophisticated (and presumably more efficient) alternative to mutations for the purpose of maintaining genetic diversity. In sexual species, each individual owns a diploid genome (consisting of two different copies). During reproduction, the parental genomes are recombined on the basis of entire chromosomes and fragments of chromosomes ("crossing over"). In contrast to mutations, this almost always leads to offspring that can survive.

Sexual reproduction

In the course of evolution, populations of organisms often separate (e.g., spatially) and develop over time into distinct species. While the differences are the result of accumulating mutations, the genomes of the descendant species still share significant similarity. In particular, many encoded proteins remain similar; such proteins are said to be *orthologs*. If proteins within the same genome are similar due to a common origin (as the result of duplications), they are called *paralogs*. *Homology* refers to any kind of evolutionary relatedness, be it orthologous or paralogous. Homologous proteins must be distinguished from *analogous* proteins, which have the same function but have evolved independently (*convergent evolution*).

1.2 Molecular Biology Measurement Data

Modern molecular biology is characterized by the (usually highly automated) collection of large volumes of data. A large number of existing measurement technologies serve to produce data on various aspects of cells and organisms. Table 1.4 provides an overview of the most common data types.

For many molecular biology data types, more than one measurement technology exists. Serious analysis must be performed bearing this in mind. For example, protein structures can be resolved either by NMR (nuclear magnetic resonance) or by x-ray crystallography. (Both methods are sciences in themselves and are hard to apply, if they work at all for a particular protein.) For some analyses, the source of the data can make a difference: apart from the lower resolution of the NMR structures, the structures may show systematic differences in the amino acids on the protein surface, because they are in contact with water for NMR whereas they are in contact with a neighboring protein in the crystal used for x-ray diffraction.

Data Type and Details	**Representation**
Sequences	
■ **DNA**: genome (hereditary information)	string over nucleotides $\{A,C,G,T\}$
■ **full-length mRNAs**: spliced gene copies	string over ribonucleotides $\{A,C,G,U\}$
■ **ESTs** (expressed sequence tags): partial mRNAs	string over ribonucleotides $\{A,C,G,U\}$
■ **proteins**	string over amino acids (size 20)
Structures	
■ **metabolites**: positions and bonds of atoms	labeled graph embedded in 3D space
■ **macromolecules** (proteins, RNAs, DNA)	labeled graph embedded in 3D space
Interactions	
■ **proteins with metabolites**: receptors or enzymes binding ligands	real vectors (binding energies)
■ **proteins with DNA**: transcription factors, etc.	binary (bipartite graph)
■ **proteins with proteins**: complexes, etc.	binary (graph); Petri-net
Expression / localization data	
■ **gene expression**: abundances of mRNAs	real vectors or matrices
■ **protein expression**: abundances of proteins	real vectors or matrices
■ **metabolite (small molecule) "expression"**: concentrations of metabolites	real vectors or matrices
■ **protein localization**: compartment of presence	categorical
Cell / organism data	
■ **genotype**: single nucleotide polymorphisms	vector of nucleotides $\{A,C,G,T\}$
■ **phenotype**: cell type, size, gender, eye color, etc.	vector of real and categorical attributes
■ **state/clinical data**: disease, blood sugar, etc.	vector of real and categorical attributes
■ **environment**: nutrients, temperature, etc.	vector of real and categorical attributes
Population data	
■ **linkage disequilibrium**: LOD scores	real numbers
■ **pedigrees**	certain (treelike) graphs
■ **phylogenies**: "pedigree of species"	trees or generalizations of trees
Scientific texts	
■ **texts**: articles, abstracts, webpages	natural language texts (in English)

Table 1.4 Common genomics data types and their representation for computational analysis.

The problems become worse for gene expression data, where the preprocessing is also crucial, as stressed below.

A few data types are so fundamental and frequent that we discuss them in the following sections.

1.2.1 Sequence Data

Sequencing

The classic molecular biology data type is the sequence (more precisely, the DNA sequence). The process of "measuring" the sequence of nucleotides in a piece of DNA is called *sequencing* and is presently highly automated. Still, it is far from trivial. First, the sequencing process requires a huge number of identical DNA molecules. These can be gained from a small sample (or even a single molecule) by *amplification* through the *polymerase chain reaction (PCR).* A more severe shortcoming is that only a few hundred up to about one thousand consecutive nucleotides of a piece of DNA can be determined in one run.

Nevertheless, is has become almost routine to sequence entire genomes. To that end, the DNA is first split into parts which are sequenced separately. The resulting set of sequences must be computationally assembled into the contiguous genome. Although techniques for the determination of protein sequences exist, it is nowadays common to sequence mRNAs (after first converting them to cDNA) or complete genomes, and then compute the translation products.

Sequence databases

Table 1.5 provides an overview of major sequence databases and portals on the Internet. Care is required when using these databases with machine learning methods: a major assumption of many algorithms, namely, that the data are *iid* (independent and identically distributed), is violated in most databases. The reason is that the proteins considered most interesting have been studied in great detail (i.e., in many species and versions) by biologists, and are therefore overrepresented. A common solution to this is redundancy reduction (the elimination of similar sequences), as provided, for instance, by the ASTRAL server at `http://astral.stanford.edu/`.

Apart from the big general sequence databases there exist a large number of more specialized databases, which often provide additional information that may be linked to the sequences. They cannot be listed here, but the journal *Nucleic Acids Research* provides reports on many of them in the first issue of each year. In addition, many of these databases are accessible via SRS.

Alignments A most basic and most important bioinformatics task is to find the set of homologs for a given sequence. Since sequences are a very important data type (not only for bioinformatics but also in other areas), new methods for sequence comparison are being developed. The new string kernels presented in chapters **??** and **??** are examples of such techniques. Nevertheless, the established methods continue to be widely used, and often serve as a crucial ingredient to machine learning approaches (cf. chapters **??** and **??**). Here we briefly introduce the the most fundamental sequence analysis technique: the alignment.

In a *global alignment* of two sequences $s = s_1 \dots s_{|s|}$ and $t = t_1 \dots t_{|t|}$, each

Database	URL (`http://...`)	Remarks
Nucleotide sequence databases		
■ DDBJ	`www.ddbj.nig.ac.jp`	these three databases synchronize their contents daily
■ EMBL	`www.ebi.ac.uk/embl/`	
■ GenBank	`www.ncbi.nlm.nih.gov`	
Protein sequence databases		
■ SwissProt	`www.expasy.org/sprot/`	curated
■ TrEMBL	`www.expasy.org/sprot/`	not curated
(Some) Sequence motif databases		
■ eMotif	`motif.stanford.edu/emotif/`	protein regular expression patterns
■ SMART	`smart.embl-heidelberg.de/`	protein domain HMMs
■ TRANSFAC	`transfac.gbf.de/TRANSFAC/`	genomic TF binding sites
General portals		
■ EBI	`www.ebi.ac.uk`	European Bioinformatics Institute
■ Entrez	`www.ncbi.nlm.nih.gov/Entrez/`	U.S. National Bioinformatics Institute
■ ExPASy	`www.expasy.org`	Expert Protein Analysis System
■ SRS	`srs.ebi.ac.uk`	Sequence Retrieval System

Table 1.5 Databases of molecular biological sequences.

Alignment

sequence may be elongated by inserting copies of a special symbol (the dash, “-”) at any position, yielding two stuffed sequences s' and t'. The first requirement is that the stuffed sequences have the same length. This allows them to be written on top of each other, so that each symbol of s is either mapped to a symbol of t (*substitution*), or mapped to a dash (*gap*), and vice versa. The second requirement for a valid alignment is that no dash be mapped to a dash, which restricts the length of any global alignment to a maximum of $|s| + |t|$. In a *local alignment*, a substring of s is globally aligned to a substring of t.

Optimal alignment

For aligning biological sequences, scores reflecting the probabilities of insertions/deletions and of mutations are assigned to gaps and to all different possible substitutions. The score of an entire alignment is defined as the sum of the individual scores. The similarity of s and t is often defined as the score of an optimal local alignment of s and t, where optimal means maximizing the score. Although there are exponentially many possible alignments (whether local or global), the optimal cost and an optimal alignment (of either mode) can be computed in time $\mathcal{O}(|s||t|)$ using dynamic programming (Needleman and Wunsch, 1970; Smith and Waterman, 1981).

Fast heuristics

Since quadratic time is still too slow for searching large databases, fast heuristics have been developed. FASTA and BLAST are much faster than dynamic programming, but achieve results almost as good. However, a much better measure of similarity can be computed by taking into account the distribution of closely related sequences. PSI-BLAST (Altschul et al., 1997) constructs a *multiple alignment* for each query sequence that consists of all similar sequences found in the database. From this a *position-specific scoring matrix (PSSM)* is constructed with which the database is searched again, thereby increasing the sensitivity of the search. This has become the most widely used method for making use of unlabeled data in supervised problems like protein classification.

Probabilistic similarity measures

With either alignment method, the obtained score depends on the length of the two sequences. For local alignments, this is compensated for by the use of so-called *p-values* or *E-values*, which quantify the chance of finding a random similarity in terms of probabilities or expected numbers of hits, respectively. Other methods of obtaining probabilistic similarity measures are based on *hidden Markov models (HMMs)* and Bayesian reasoning. An excellent textbook on alignments and related topics is Durbin et al. (1998).

1.2.2 Gene Expression Data

Gene expression data usually come in the form of a matrix of expression levels for a number of genes in a range of cell samples. There are quite a few technologies available to measure the level of expression of a large number of genes simultaneously. We focus on microarrays, which are presently the most popular technology for large-scale gene expression measurement. Then we outline two competing techniques that are based on a different approach. Finally, we discuss some implications for data analysis.

Microarrays

Microarrays Microarrays, sometimes also called DNA chips, employ hybridization to distinguish different genes, and therefore require that the sequences of genes to be measured be known in advance. In fact, a *microarray* is essentially a surface with a known location (called *spot*) for each gene to be measured. Present-day microarrays can bear a couple of thousand spots, so that the entire human genome can be covered with four chips. At each spot, oligonucleotides or cDNA fragments are fixed which are complementary to a (transcribed) subsequence of a gene. Ideally, the subsequences are determined in such a way that they are specific to the corresponding gene, that is, they are not similar to the complement of any other mRNA that is expected to occur in the sample.

Measurement (hybridization)

The measurement of a sample with a microarray (jargon: *hybridization*) begins by reverse-transcribing the mRNAs of a cell sample to cDNA. The cDNAs are labeled to make them detectable, for instance, by incorporating fluorescing or radioactive tags. Then, the sample is administered onto the microarray, and a number of cDNAs from the sample hybridize to the corresponding spot. This number is approximately proportional to the respective mRNA concentration in the sample. After washing the array the concentration can be determined by measuring, at the corresponding spot, the intensity of the signal emitted by the molecular labels. If two different molecular labelings are used for two different samples, two measurements can be carried out at the same time on the same array.

Noise

A couple of facts are essential to the proper analysis of microarray data. First, *background noise* is present due to incomplete washing and nonspecific hybridization. If the mean background is estimated and subtracted, the resulting expression levels may become negative for some genes. Although true expression levels cannot be negative, statistical work seems to suggest that it is not a good idea to censor such values by setting them to zero or a small positive value; instead, variance-stabilizing transformations may be used. Second, due to varying efficiencies of the intermediary steps of the measurement, and to varying amounts of mRNA per cell, the results obtained with different microarrays or for different samples are not likely to be on the same scale. *Normalization* should therefore be applied. Moreover, systematic differences that arise from fluctuations between production batches should be compensated for. Finally, even normalization cannot make comparable data gathered with microarrays that are equipped with different oligonucleotides (e.g., chips of different brands). This is because the oligonucleotides have different hybridization energies which introduce a scaling constant for every spot (gene).

Normalization

Other technologies A few methods for measuring gene expression levels are based on sequencing, clustering, and counting mRNA molecules, as detailed in the following three-step strategy:

1. *Sequencing.* This involves randomly picking mRNA molecules from the sample, reverse-transcribing them to cDNA, amplifying them, and then determining (parts of) their sequences.
2. *Clustering.* Sequences corresponding to the same genes must be identified.
3. *Counting.* The cluster sizes are estimates of the expression levels.

Two examples from this group of methods are *serial analysis of gene expression (SAGE)* and *expressed sequence tag (EST)* analysis. An EST results from partial single-pass sequencing of a transcript; it represents an error-prone substring of a (usually spliced) mRNA.

ESTs, SAGE

In contrast to microarray measurements, counting sequence tags (as in SAGE or EST analysis) yields natural numbers as outputs. The difficulty of analysis arises from three facts: (1) The number of sampled molecules is low (ESTs) to medium (SAGE) due to the effort (and cost) required. Thus, the counts are bad estimates of the true frequencies, especially for the low copy genes, which may not be detected at all. (2) Sequencing errors may lead to inclusion of a sequence in the wrong cluster. (3) Clusters may erroneously be merged (e.g., for very homologous genes) or split (e.g., if sequenced parts do not overlap). The importance of sequencing-based methods is that they do not require prior knowledge of the genes; therefore, they can in principle be applied to any genome.

Two other methods deserve mention, because they are frequently used by biologists. *Northern blotting* is the oldest approach to (semiquantitative) measurement of gene expression. In contrast, *quantitative polymerase chain reaction (qPCR)* is a very modern method, and is currently considered to allow for the most precise determination of expression levels. Both methods require prior knowledge of the

Northern blotting, quantitative PCR

sequence. In addition, they are not sufficiently automated for mass measurements; instead, they are used to confirm findings made with other technologies.

Databases Unfortunately, the databases for gene expression data are not yet as established as the sequence databases are. It is still common for microarray data sets to be available only from webpages that accompany a publication. Nevertheless, there are two databases that try to be very general. These and a few databases with more specialized domains are listed in table 1.6.

Database	**URL** (`http://...`)	**Remarks**
General databases		
■ ArrayExpress	`www.ebi.ac.uk/arrayexpress/`	by the EBI
■ GEO	`www.ncbi.nlm.nih.gov/geo/`	by the NCBI
Organism-specific databases		
■ MGI GXD	`www.informatics.jax.org`	mouse
■ TAIR Microarray	`www.arabidopsis.org`	*Arabidopsis*
■ WormBase	`www.wormbase.org`	*Caenorhabditis elegans*
Laboratory-specific databases		
■ SMD	`genome-www.stanford.edu/microarray/`	Stanford
■ YMD	`info.med.yale.edu/microarray/`	Yale

Table 1.6 Major databases of gene expression data sets.

One reason for the slow establishment of general gene expression databases may be that it is surprisingly difficult to properly design its scheme. To be really useful, there is no point in just storing the matrices of measurement values. Instead, a description of the preprocessing of the data, the measurement technology (including details of the biochemical steps carried out in the laboratory), and above all the properties of the samples, must be given. For samples taken from hospital patients, for example, a complete description would ideally include the entire clinical data.

1.2.3 Protein Data

Reflecting their importance for cells, many aspects of proteins other than their sequences are wellstudied, including (tertiary) structures, interactions, functions, expression, and localization. Many of these data may be found in databases; see table 1.7.

The unique worldwide database of protein structures is the PDB (protein database). However, detailed structure comparisons are tedious, and thankfully databases exist that provide hierarchical classifications of the PDB proteins (or the domains therein) according to their structures. The most popular may be SCOP (“structural classification of proteins”, which is largely manually constructed by ex-

Database	URL (http://...)	Remarks
Protein structures		
■ PDB	www.rcsb.org/pdb/	3D structures
■ SCOP	scop.mrc-lmb.cam.ac.uk/scop/	structural classification
■ CATH	www.biochem.ucl.ac.uk/bsm/cath/	structural classification
Molecular interactions and networks		
■ BIND	www.bind.ca	interaction network
■ KEGG	www.genome.ad.jp/kegg/	metabolic pathways
■ DIP	dip.doe-mbi.ucla.edu	interacting proteins
Protein functions		
■ GO	www.geneontology.org	controlled vocabulary
■ EC	www.chem.qmul.ac.uk/iubmb/enzyme/	enzyme numbers
■ MIPS	mips.gsf.de/proj/yeast/ catalogs/funcat/	yeast gene functions
Protein expression		
■ 2DPAGE	us.expasy.org/ch2d/	2D gel electrophoresis data

Table 1.7 Important databases on protein properties other than sequence.

perts) and CATH ("Class, Architecture, Topology and Homologous superfamily", which relies more heavily on automatic classification).

Protein *interactions* can be defined on several levels: *molecular interactions* refer to the binding partners of proteins, while the broader notion of *regulatory interactions* also includes indirect influences like up- or downregulation through signaling pathways. Proteins can also be linked in a *metabolic pathway* if they catalyze successive steps in a series of metabolic reactions. Any such interactions, also with other molecules like DNA, can be used as edges in a *biological network* graph. A couple of databases provide interactions measured or inferred by different methods.

There are also databases providing functional classifications of proteins (or of the respective genes). Protein function is related to interactions, localization, and expression. There are two popular ways to measure protein expression: (1) by *2D gel electrophoresis*, in which proteins are spatially separated in a gel according to mass and electric charge; and (2) by *mass spectrometry (MS)*, in which the masses of protein fragments are very precisely inferred by measuring their time of flight after a defined acceleration. However, to the best of our knowledge, no general databases containing MS protein expression levels exist yet; nor do such databases exist for protein localization.

1.2.4 Other Data Types

A number of data types additional to those discussed above also deserve mention. (1) Chemical compounds (small molecules) are interesting as metabolites, signaling molecules, and potential drugs. (2) SNPs account for most phenotypic differences between individuals, including susceptibility to many diseases. (3) The abstracts of the scientific molecular biology publications form a huge reservoir of badly structured data. These are the targets of text mining, which attempts to automatically extract information. Sources of these three types of data are listed in table 1.8; in addition, the URLs of two database directories are given in table 1.9.

Database	**URL** (`http://...`)	**Remarks**
dbSNP	`www.ncbi.nlm.nih.gov/SNP/`	single nucleotide polymorphisms
PubMed	`www.ncbi.nlm.nih.gov/PubMed/`	publication abstracts
NCI	`cactus.nci.nih.gov`	small molecule structures

Table 1.8 Databases on data types not covered in the previous tables.

URL (`http://...`)
`bip.weizmann.ac.il/mb/molecular_biol_databases.html`
`molbio.info.nih.gov/molbio/db.html`

Table 1.9 Collections of links to databases on the web.

Model organisms

Model organisms are organisms chosen by biologists to be representative of some class or property and, at the same time, to be simple and accessible. For example, the fruit fly *Drosophila melanogaster* shares many genes and somatic functions with humans, but is much easier to investigate (no ethical problems, short reproduction cycle, etc.). The value of model organisms for bioinformatics lies in the fact that not only are (almost) complete genomes available for them (this is now the case for hundreds of organisms) but also plenty of other data which can be set into relation with the genes. Some model organisms are compiled in table 1.10.

We conclude this section with a small disclaimer: our intention here is to provide pointers to the largest and most general mainstream databases of bioinformatics. However, there exist a large number of additional databases on various (often rather specialized) molecular biological problems. Once again, we refer the interested reader to the annual database issue of *Nucleic Acids Research*.

1.3 Bioinformatics Challenges

As described in section 1.2, modern molecular biologists measure huge amounts of data of various types. The intention is to use these data to (1) reconstruct the past (e.g., infer the evolution of species); (2) predict the future (e.g., predict how someone will respond to a certain drug); (3) guide biological engineering (such

Organism (Common Name)	**Level**	**Cells**	**Genes**	**Genome**
Human immunodeficiency virus (HIV)	virus	0	9	0.01 Mb
Methanococcus jannaschii	archaea	1	1750	1.66 Mb
Escherichia coli (human gut bacteria)	eubacteria	1	4300	4.6 Mb
Saccharomyces cerevisiae (brewer's yeast)	eukaryote	1	6000	12 Mb
Caenorhabditis elegans (nematode worm)	animal	959	19,500	100 Mb
Drosophila melanogaster (fruit fly)	animal	?	13,700	165 Mb
Arabidopsis thaliana (thale cress)	plant	?	25,498	125 Mb
Mus musculus (mouse)	mammal	?	35,000	3000 Mb

Table 1.10 A selection of model organisms. We were unable to obtain information on the number of cells in the last three organisms; however, it is estimated that an adult human body consists of about $6 \cdot 10^{10}$ cells. The numbers of genes are estimates (except for HIV). The genome size is given as the number of nucleotides in a single strand of the haploid genome (Mb, for 10^6 basepairs).

as improving the efficiency of brewer's yeast). Some of the concrete tasks are so complex that intermediate steps are already regarded as problems in their own right. For example, while the sequence of a protein in principle determines its function (in the particular environment provided by the cell that produces the protein), one of the grand challenges in bioinformatics is to predict its structure (which can then serve as a basis for investigating functional aspects like interactions). This can also be seen as an auxiliary goal complementing the three listed above: to replace difficult, laborious, time-consuming, and expensive measurements (x-ray crystallography) with more affordable ones (sequencing).

The rather vague goals described above manifest in a jungle of concrete computational problems. Some of them are very specific, for instance, related to a certain species or a particular protein; their solution can aid in a particular application. Many other problems are more fundamental, but often can be seen as building blocks for the solution of larger tasks. In the following subsections, we try to organize some of the more fundamental challenges into a small number of general schemes. To some of the problems described below, existing approaches are reviewed in chapter **??**.

The following sections are organized along the drug development process of pharmaceutical companies, which is one of the main applications:

1. Understanding the biological system, especially the mechanism of the disease
2. Identifying target proteins which are pivotal for the disease
3. Characterizing those proteins; most importantly, their tertiary structure
4. Finding small molecules which bind to those proteins and which qualify as drugs

1.3.1 Genome Structure Analysis

The analysis of DNA sequences (partial or complete genomes) can be organized into a small tree, as depicted in figure 1.7. It contains at least three “grand challenge” problems:

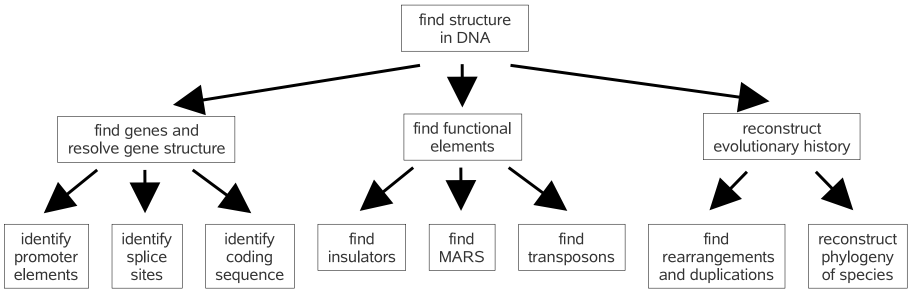
find structure in DNA
find genes and resolve gene structure
find functional elements
reconstruct evolutionary history
identify promoter elements
identify splice sites
identify coding sequence
find insulators
find MARS
find transposons
find rearrangements and duplications
reconstruct phylogeny of species

Figure 1.7 Finding structure in genomic DNA sequences: a small hierarchy of problems.

- *Genome comparison.* The goal of this discipline is to reconstruct the evolutionary history, that is, the series of genome rearrangements, that led to different species. A difficulty is that the phylogeny and the common ancestors must be inferred on the basis of genomes of present-day species. While pairwise whole genome comparisons are already a challenge due to the sheer size of a genome, only comparison of multiple genomes will unleash the full power of this approach. Here, most efficient algorithms are asked for.
- *Gene finding.* This includes the identification of the gene structure, that is, the arrangement of the gene’s elements (introns, exons, promoter, etc.). In computer science terms, the problem is to label substrings of the DNA. In large genomes with low gene content, like the human genome, especially the false positives can be a problem. An accurate solution to a large subproblem of gene structure identification, the prediction of splice sites, is described in chapter **??**.
- *Understanding transcriptional regulation.* Here the goal is to quantitatively predict the expression levels of genes from the details of their promoters and the present quantities of TFs. In its broadest sense this problem would also include *modeling the 3D structure of DNA*. The packing of DNA is believed to have a big impact on gene expression, since genes must be unpacked before they can be transcribed. However, the available data may not yet be sufficient for such modeling.

All tasks are complicated by the fact that (presumably) by far not all functional DNA motifs are known by now. In fact, the understanding of the huge part of DNA which cannot yet be assigned a function can also be seen as a grand challenge, albeit possibly in molecular biology rather than in bioinformatics.

1.3.2 Relation of Molecular to Macroscopic Data

It is most interesting to identify the molecular causes of macroscopic events or states, because this understanding allows for a directed search for ways to cause or prevent such events and to maintain or change such states. Figure 1.8 provides examples of the molecular and macroscopic data types to be causally related. Three classes of tasks emerge from this general problem statement:

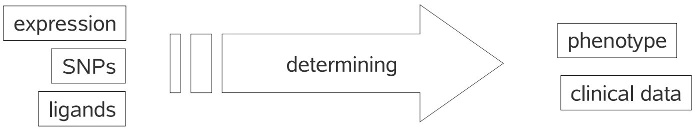

Figure 1.8 Macroscopic states (*right*) are caused by molecules (*left*); thus, molecular measurement data contain information predictive of the macroscopic states.

- *Population genetics.* The strategy in population genetics is to find chromosomal regions that are inherited along with (completely or partially) heritable traits; such regions can be supposed to contain genes responsible for those traits. The basic data for this are pedigrees of families which are annotated with both phenotypic and genotypic information on the individuals. The genotypic part consists of so-called *genetic markers* (such as SNPs) that relate genetic content to chromosomal location.
- *Diagnosis.* As an example, it is desirable to be able to base the diagnosis of certain diseases on gene expression patterns. For diseases that are hard to recognize or distinguish by classic means (e.g., histology), this can potentially be less subjective and ambiguous. Genetic diseases may also be diagnosed based on SNPs, or both SNPs and expression data. Sometimes unsupervised analysis of molecular data can even lead to refined definitions of diseases (Golub et al., 1999).
- *Therapy optimization.* Here, the idea is that every individual is different and that optimal treatment may be derived from molecular data: the efficacy of drugs may be predicted on the basis of the genotype of a pathogen (Beerenwinkel et al., 2003). Optimally, the interplay of the genotype (SNPs) of the patient with that of the pathogen should be taken into account.
- *Target finding.* This essentially amounts to applying feature selection to a successful prediction of a disease (diagnosis); see chapter **??**. Ideally, the relevant features are related to the cause of the disease, which can then be selected as the target of drug development.
- *Systems biology.* This is the most ambitious challenge under this rubric, and is likely to keep bioinformaticians busy in coming years: the goal is nothing less than to quantitatively simulate entire cells (or large subsystems). This would (among

many other things) allow replacement of animal experiments by computational simulations. A step in that direction has already been taken by E-CELL (Tomita, 2001).

1.3.3 Protein Property Prediction

Trying to predict properties of proteins is alluring, because proteins are so important for the cell (and for the biologist), and difficult, because proteins are so complex and versatile. There is a whole family of problems which are distinguished by the predicted property and by the data on which the prediction is based; this is sketched in figure 1.9.

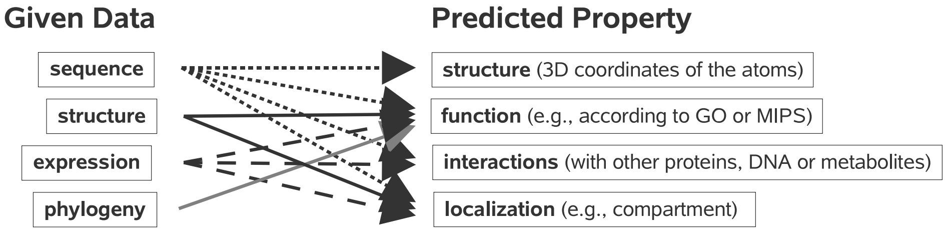
Given Data
sequence
structure
expression
phylogeny
Predicted Property
structure (3D coordinates of the atoms)
function (e.g., according to GO or MIPS)
interactions (with other proteins, DNA or metabolites)
localization (e.g., compartment)

Figure 1.9 The prediction of different properties of proteins can be based on different (combinations of) data types.

There are also a number of prediction problems that are not explicitly shown, because they can be seen as intermediate steps. They include the prediction of structural motifs (most important, secondary structure) and of solvent accessibility of amino acids, which can provide valuable hints for a structure prediction. Another analytical task is the prediction of modifications (such as phosphorylation) from sequence. Modifications can affect all four types of properties shown on the right side of figure 1.9.

At least four grand challenge problems are instances of the family illustrated in the figure:

- *Structure prediction.* Here the amino acid sequence is given, and the 3D structure of the folded protein (in a cell) is to be computed. The fact that the cellular environment is so important in reaching the correct structure (cf. section 1.1.3) renders the idea of simulating its molecular motion in watery solution unappealing (although this is tried with vigor). Indeed, the most successful structure prediction methods are knowledge-based (i.e., at least in a way, machine learning) methods.[5]

[5] This is demonstrated in the evaluation of the CASP competition available from

Chapters **??** and **??** describe applications of kernel methods to structure prediction.

- *Function prediction.* Here the problem starts with finding an appropriate definition of function. While hierarchical classifications of functions now exist, it is not clear whether they are well suited to serve as classes to be predicted. (At least, the idea of cutting the hierarchy at a fixed depth to define the classes seems to be questionable.) It has been shown that the best performance can be achieved by making use of multiple data types, although the proper combination is not trivial (Pavlidis et al., 2002, see also chapters **??** and **??**).

- *Genetic network reconstruction.* The term *genetic network* refers to a graph specifying interactions between molecules, especially of regulatory type. Although a few experimental methods exist to find such interactions, there is also great interest in predicting them to obtain a more complete picture at less cost. A genetic network can allow deduction of hypotheses about, say, the effects of inhibiting certain proteins, which may suggest drug target candidates. Several models for computational treatment have been suggested: probably the most prominent classes are Boolean, linear, and Bayesian networks; a recent trend is the use of graph kernels (cf. chapter **??**).

- *Docking.* This is the computational prediction of molecular binding events involving proteins. There are two flavors of protein docking: protein-protein docking and protein-ligand docking. The goal of protein-protein docking is basically to predict whether two proteins can bind at all; this can contribute edges for biological networks. While the backbones are usually taken to be fixed, it is important to model the flexibility of the involved side chains. In protein-ligand docking, the flexibility of the *ligand* (the small molecule) is essential; often, it is also crucial on the side of the protein (*induced fit*). Here, the goal includes predicting the strength of the binding (*affinity*), which must be high for inhibitors.

1.3.4 Small Molecule Problems

Chemoinformatics, cheminformatics

Computational work with chemical compounds is sometimes regarded as a subdiscipline of bioinformatics and sometimes viewed as a separate field termed *chemoinformatics* or *cheminformatics.* For completeness, we briefly introduce the main problems in this area that bear biological relevance. Usually they are closely related to the task of drug development.

- *Virtual HTS (high-throughput screening).* This is essentially protein-ligand docking (see above) viewed from a different perspective. Simulating the experimental HTS, large databases of compounds are tested against a receptor protein to identify potential ligands.

- *Lead identification.* This means proposing a novel skeleton of a suitable drug molecule, called a *lead structure*, based on a collection of data related to a disease.

http://predictioncenter.llnl.gov/casp5/.

The data may be the product of virtual or real HTS. While virtual HTS relies on the atomic structure of the binding pocket of the protein of interest, in practice it often is not known. Then techniques like *QSAR* (quantitative structure-activity relationships) are used to infer important properties from molecules with known activity. Such properties can be summarized in a *pharmacophore*, an abstract characterization of the set of molecules of interest. In *lead hopping*, the task is to find new lead structures for a given pharmacophore (with known leads), for example, to evade patent problems.

■ *Predictive toxicology.* Not being (too) toxic is an important prerequisite for a drug. Depending on the time scale of the effect, toxicity may be acute or chronic. While acute cytostatic (inhibiting cell growth) toxicity is quite amenable to experimental investigations, lab screenings for chronic effects like carcinogenicity (causing cancer) are very time-consuming and hard to standardize. Thus, reliable in silico predictions would be of high value. Of course, there are further properties of small molecules that are relevant to drug candidates, often subsumed by the term *ADME* (*a*bsorption, *d*istribution, *m*etabolism, and *e*xcretion). Much effort has been spent on developing good representations of molecules for solving such prediction tasks; modern methods allow working directly with the natural representation by a labeled graph (cf. chapter **??**).

1.4 Summary

It is hard to summarize the first two sections of this chapter. The section on the basic biology of the cell (section 1.1) is already a summary in itself. Please be warned that there is much, much more to cellular biology and that it really matters for successful bioinformatics. With respect to the molecular biology measurement data (section 1.2), we would like to add that biotechnology is a very active field and new technology is constantly being developed. Thus, new exciting types of data can be expected to become available and popular in the near future.

What we do want to explicitly point out here are a couple of things that may have already become apparent from section 1.3. In several examples it could be seen that many bioinformatics problems

1. can be posed as machine learning problems;
2. concern a structured data type (as opposed to “simple” real vectors);
3. concern a combination of different data types.

Corresponding to that, the subsequent chapters of this book are concerned with the following questions:

1. How to properly model bioinformatics problems in a machine learning framework.
2. How to operate on structured data types with the use of kernel functions.

3. How to combine different data types with clever kernels or algorithms.

While there already is some tradition of machine learning approaches to the analysis of molecular biology data, it has so far been mostly concerned with relatively simple data structures (usually, fixed-length vectors of real values and categorical attributes). By demonstrating new ways to deal with complex data this book may contribute to accelerating the progress and success of machine learning in bioinformatics, and possibly also in other application areas.

References

B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell.* New York, Garland Science Publishing, 1998.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* New York, Garland Science Publishing, 4th edition, 2002.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

N. Beerenwinkel, M. Daumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, L. Lengauer, J. Selbig, and H. Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Research*, 31(13):3850–3855, 2003.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, UK, Cambridge University Press, 1998.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

P. Pavlidis, J. Weston, J. Cai, and W. S. Noble. Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9(2):401–411, 2002.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

M. Tomita. Whole-cell simulation: A grand challenge of the 21st century. *Trends in Biochemical Sciences*, 19(6):205–210, 2001.